AUTOMATIC DETECTION OF LIVER TUMORS

Daniel Pescia[1,2,3] *Nikos Paragios*[1,2] *Stéphane Chemouny* [3]

(1) Laboratoire Mathématiques Appliquées aux Systèmes, Ecole Centrale de Paris
(2) Equipe GALEN, INRIA Saclay, Ile-de-France
(3) Intrasense SAS, Montpellier, France

ABSTRACT

Tumor detection in CT liver images is a challenging task. The nature of tumor has a direct effect on the number of voxels being contaminated, as well as on the changes in the observed CT scan. In order to deal with this challenge, in this paper we propose the use of advanced non-linear machine learning techniques to determine the optimal features, as well as the hyperplane that use these features to separate tumoral voxels from voxels corresponding to healthy tissues. Very promising classification results using an important volume of clinically annotated data (86% sensitivity, 82% specificity) demonstrate the potentials of our approach.

***Index Terms*—** Image segmentation, Liver tumors, Machine Learning, Texture, AdaBoost

1. INTRODUCTION

Liver cancer is a deadly disease with an important frequency on the world. Surgical resection is the best treatment available, but may apply only when some conditions on tumor sizes are met. Therefore early diagnosis and accurate appraisal of tumors are critical. The exceptional resolution of CT images allows good detection rates for most tumor types. However, the detection of liver tumors is challenging due to the small observable changes between healthy tissues and tumoral ones. Such a task is even challenging for clinical experts, where one can refer to an important volume variation (15-25%) between experts. Thus a good and repeatable method to detect the tumors would be an advantage. Being clinically motivated, such an approach should work for diverse tumor types at same time, in order to avoid multiple and successive segmentations.

Such a task is quite difficult. First, various tumors have to be segmented, with miscellaneous appearances. Then, often a striking resemblance exists between tumoral and healthy tissues. Moreover, some tumors are only visible, or more distinguishable, when an appropriate phase has been considered. In general, these phases are roughly four, which correspond to diverse delays between the injection of a contrast product and the image acquisition. At the image level, the outcomes are diverse appearances and intensities ranges, both for the healthy tissues and the tumoral ones, along with shifting appearances for a same tumor from one phase to another. Another challenging problem is due to the liver itself. Being an organ with a high level of vascularization, the images obtained are really noisy, a property that is amplified due to the enhancements. Finally, because of the focus on clinical use, the method should work for real images, meaning images from diverse CT machines, with various levels of resolution and for both connected as well as disconnected slices.

In this paper, we would like to address the problem of classification of diverse tumor types versus the healthy tissues in the liver. This detection has to be done towards satisfying a number of constraints, like different resolution levels, various enhancement phases and protocols, and in noisy anisotropic CT images. We propose a learning based approach to address the task, which address both feature selection and tumor classification. We assume that image intensities are normalized, and from a predefined set of filters we select the ones for which an optimal classifier exists, to separate healthy versus non-healthy tissues. This classifier is based on the AdaBoost method [1], that decomposes the process into a number of weak classification tests. Once such a multi-level, multi-feature classifier has been determined, the task of detection consists of determining the distance between the classifier and the observations in a new volume.

The remainder of the paper is organized as follows: in section 2 we briefly present the feature bank and the classifier while the task of detection is part of section 3. Experimental results and conclusions are part of section 4.

2. BACKGROUND

Machine learning aims to determine a process that separates a set of observations. In our case, observations consist of intensities and classes correspond to tumor versus non-infected tissue. Despite the resolution of CT images, one can imagine that the separation of healthy versus non-healthy samples in this space is almost impossible. The use of filters and their responses, is a convenient way to take into account the relative context, and consider features with better discrimination

This work was partially supported by ANR grant SURF-NT05-2_45825.

power. One can consider either the responses themselves or seek for a separation on a subspace that encodes the dependencies at the local scale of these responses.

First, let us consider without loss of generality a bank of texture descriptors. Then, let us use the AdaBoost method to learn a separation in the feature space.

2.1. Texture descriptors

One can seek for a feature space that measures the statistical properties of this space. We have considered two different texture descriptors. First, statistical metrics on the texture histogram were chosen (standard deviation, skewness, etc). Then, some second order descriptors were retained, namely Haralick's descriptors [2], because they have been shown to be the most informative ones in a similar context [3].

These Haralick's descriptors are metrics upon pairs of pixels. They are computed using co-occurrence matrices $P(d, \theta)$, which are distribution of probabilities, that keep track of pixels pairs for given direction θ and distance d. $P_{i,j}(d, \theta)$ gives the probability to get a pair of pixels with intensities of i and j, at distance d and in direction θ. Because of the wide range of intensities, these matrices may have a high size, without information gain, thus intensities are linearly normalized on each texture. It allows to narrow the dimension of the co-occurrence matrices to an admitted number of gray levels, and with a substantive information gain.

Once the feature space has been determined, a classifier is to be introduced towards separating the features.

2.2. AdaBoost

AdaBoost is a supervised learning method, introduced by Freund and Schapire[1], based on the use of weak learners to construct a strong classifier. This method has been widely used, because it runs fast (when the weak learners are fast) and may be applied in many cases. However, the quality of the results is dependent of the choice of adapted weak learners. This method applies well to our problem, because it allows to account for classes that may be divided into clusters.

AdaBoost uses a training set (1), and a set of weak learners to learn a strong classifier (2). The training set is made up of pairs of object or features representing the object, noted x_i, and related class, noted y_i. A strong classifier is created by an an iterative process, as a linear sum of weak learners. This gives a classification function for the training set, thus, this last should be a good sample of the general case, in order to achieve a good generalization error.

$$\{(x_1, y_1), \ldots (x_m, y_m)\} \tag{1}$$

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{2}$$

3. METHOD

3.1. Normalization

The diversity of the exams used prevents direct comparison of the images. Thus, images are normalized by equalizing the histograms in the liver envelope.

3.2. Texture features

Instead of working on texture patches, the classification is done on the texture features of each patch. These features are obtained by cross-product between a set of filters and a set of texture descriptors. It means that each feature is a texture descriptor applied to a texture patch from a filtered image.

The use of filters is consistent with the framework. First, we work on noisy images, thus filtering is required. Then, 3D filters may have a normalization effect between the various sizes of slice. Finally, filters may be used to enhance some features. The usual filters (Gaussian, Mean, etc) are used, with diverse radius. In addition, 3D Gabor's filters were chosen to consider the texture information. Gabor's filters are strongly related with the human visual system, and are often linked with texture constraints. Finally, useful filters in the liver case were retained, namely median and Nagao's filters [4]. In the following parts, filters will be noted $f_{m,\Theta}$, with m defining the type of the filter, and Θ its parameters.

Statistical and Haralick's descriptors are used as texture descriptors, but with some refinements. The idea is to introduce some kind of multi scale approach at the texture level, meaning to use diverse sizes of texture at same time. Instead of manipulating a set of texture with diverse size, the texture descriptors were modified to account for diverse sizes.

In terms of statistical descriptors, one can account for the above modification through a histogram computation within a radius from the center. For Haralick's descriptors the change is made in the co-occurrence matrices by adding a radius r to their definition, that becomes $P(d, \theta, r)$. This new matrix gives the probability of pixels pairs at distance d, for direction θ and within a distance r from the texture center. This radius r may take any value between 1 and the texture radius. A radius of 0 make no sense, because pixels pairs are needed, and we should remain within the texture patch. The other parameters, d and θ, keep usual values, namely $d = 1$ and $\theta = \{0, 45, 90, 135\}$. It should be noted that the computation of a co-occurrence matrix $P(d, \theta, r + 1)$ is eased when $P(d, \theta, r)$ is known. One only has to add pairs on the edges.

3.3. Learning Step

First, weak learners are defined. In our case, texture features contain the information, thus the weak learners will be elementary. They are defined as the comparison of one texture feature to a reference value. For the j^{th} feature of the object

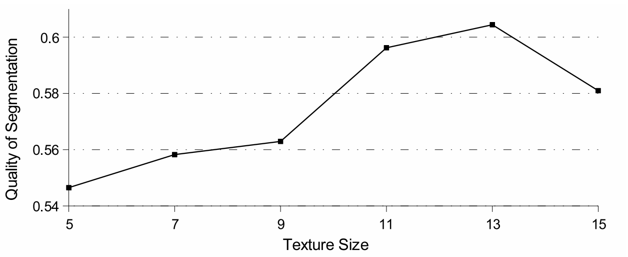

Fig. 1. Quality of segmentation in function of the texture size. An optimal size of texture of 13 is easily seen.

x a weak learner is defined as following:

$$h_l(x^j) = \begin{cases} 1 & if \;\; \delta \cdot x^j \leq \delta \cdot \gamma_{j,l} \\ -1 & otherwise \end{cases} \tag{3}$$

where $\delta \in \{-1; 1\}$ gives the sign for the comparison, and $\gamma_{j,l}$ is one of the admitted value for the j^{th} texture feature.

Then, a learning set is created. This step is truly crucial. In order to achieve good generalization error, tumoral and healthy textures should be correctly sampled, while sticking close to the real distribution of the tumor types. First, a set of exams was selected, with diverse sizes of slice and tumor types, in an attempt to sample the real distribution of the tumoral types. Then, texture patches were taken under some constraints. An equal number of tumoral and healthy patches are sampled. Theses patches are chosen with a regular distribution, in order to have a good sampling of the possible appearances. It should be noted that no texture patches were taken in the tumor edges, because the appearance here is too random and may be source of errors.

3.4. Learning Database

We have used 15 manually annotated volumes to train our classifier. These volumes include 3 types of tumors. We have also considered filters of different scales and we did compare the performance towards determining the optimal bandwidth.

4. RESULTS

Let us now consider a new volume, where one has to determine or not the presence of tumors. In order to quantify the performance of our method, one has first to define a metric with respect to the ground truth.

4.1. Comparison metric

Usually two metrics are used to quantify the quality of a segmentation, the sensitivity (4), that gives the percentage of tumor that is correctly classified, and the specificity (5), that quantifies the quality of the segmentation for healthy tissues.

	Whole Set	HCC	Metastases	Other
Sensitivity	86%	84%	87%	93%
Specificity	80%	79%	81%	84%
Specificity (vessels segmented)	82%	80%	84%	84%

Fig. 2. Segmentation results obtained with the same strong classifier for diverse tumors. Segmentation of vessel networks as a first step does not modify the sensitivity, whereas specificity improves.

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{5}$$

where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP the number of false positives.

The goal is to insure a good sensitivity without loss in term of specificity. Thus a comparison metric (6) is defined as a weighted sum of these square two terms, allowing to increase the penalty for small values. These two terms are balanced in order to emphasize the relative weight of each term. In a clinical context sensitivity matters more, thus the weights will be set to $\lambda = \frac{2}{3}$ and $\beta = \frac{1}{3}$ in the following parts.

$$\lambda \cdot [\text{sensitivity}]^2 + \beta \cdot [\text{specificity}]^2 \tag{6}$$

4.2. Impact of texture size

The choice of a size of texture is important. One has to find a balance between the computation time, that is lower for small textures, and the quality of segmentation, that should increase with texture size. The best results were obtained for a size of 13 pixels (see Fig.1). This choice is consistent with texture sizes chosen in similar problems, 9×9 windows in [3], and windows from 9×9 to 13×13 pixels in [5].

4.3. Results on diverse tumor types

Classification has been applied to a set of 798 slices, containing metastases from diverse primary sites, HCC, Adenoma and Cholangiocarcinomas. Diagnoses were confirmed by anatomical pathology. The results are quite good, with slight differences between the tumor types (see Fig.2). The high results for the less common tumor types may not be really significant, because few slices of these types were used.

The volume of the tumor impacts the quality of the segmentation, but two interesting facts should be noted. First, when tumors are bigger than the size of the texture, the detection rate is quickly optimal. Then, even tumors smaller than

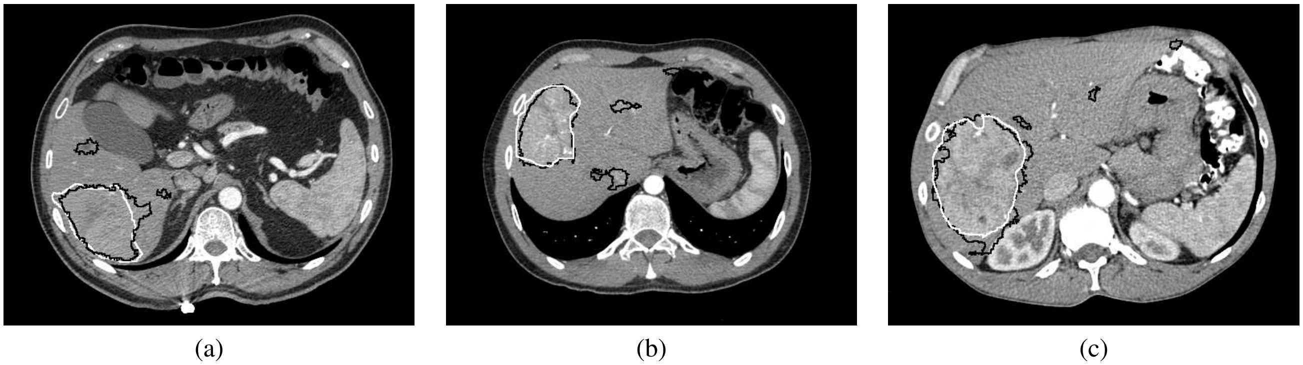

Fig. 4. Segmentation results for some tumors with a texture size of 13, without removing vessels. The ground truth is in white, the result of the segmentation is in black. From the left to the right: (a) Segmentation for metastases, portal phase, slice size 1.3mm. (b) Segmentation for a HCC, arterial phase, slice size 1.3mm. (c) Segmentation for an Adenoma, arterial phase, slice size 2mm. CT images are courtesy from V. Vilgrain's Department (Beaujon, Paris)

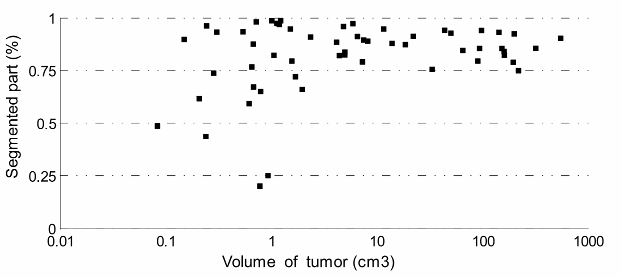

Fig. 3. Percentage, in volume, of tumors accurately segmented, in function of the volume of the tumor. The volume of the texture patches used, spans between 0.61 and 3.05 cm^3, with an average volume of 1.54 cm^3.

the patches are partially segmented, but the accuracy of the segmentation is lower and more random (see Fig.3).

There are two common cases of false positive, namely regions close to enhanced vessels, or on the sides of the liver. The first problem may be partially addressed by doing a segmentation of the liver vessels as a first step, which will help to avoid classifying vessels as tumors. This improves slightly the specificity of the classification, as may be seen in (Fig.2). The improvement remains small. Indeed this step cannot help the classification of the neighborhood of vessels, because tumors are often close to vessels and even growing new vessels.

5. CONCLUSION

We have presented a method of detection for the liver tumors. This segmentation is done by classifying pixels as their surrounding texture, with the help of an AdaBoost classifier. The results of segmentation were 86% of sensitivity and 80% of specificity. As improvement, liver vessels were removed, which turns into an increase of the specificity to 82%. It remains to be seen if this improvement is worth the time needed for vessel segmentation. In the future the classification may be improved by treating differently regions close to vessels and the boundaries of the liver. Furthermore, the use of techniques that aim to impose global consistency on the classification, using for examples MRFs or mean-shift could be a quite promising direction. Because this approach is generic enough, it may apply on other segmentation problems, where texture information is important.

6. ACKNOWLEDGEMENT

This work was made in a joint effort with Intrasense. I would like to thank Lawrence Pupulim for providing me CT images with manually segmented tumors.

7. REFERENCES

[1] R-E. Schapire Y. Freund, "A decision-theoretic generalization of on-line learning and an application to boosting.," *J. Comput. System Sci.*, 1997.

[2] I. Dinstein R. Haralick, K. Shanmugam, "Textural features for image classification.," *IEEE Trans. on SMC*, vol. 3(6), pp. 610–621, 1973.

[3] T. Disney D. Raicu J. Furst M. Pham, R. Susomboon, "A comparison of texture models for automatic liver segmentation," *SPIE Medical Imaging Conference*, 2007.

[4] S. Chemouny, "Filtrage et segmentation d'images tridimensionnelles : Application à la détection et à la caractérisation des structures anatomiques et pathologiques du foie," 2001.

[5] H. Kobatake S. Nawano L. Tesar D. Smutek, A. Shimizu, "Texture analysis of hepatocellular carcinoma and liver cysts in ct images," *SPPRA'06*, pp. 56–59, 2006.